

## Project Alexandria Repository : Underlying Philosophy and Overview of the Data Extract

---

This page last changed on Mar 04, 2009 by [davem](#).

*Copyright by Matrixware. This document is intended only to be distributed to Matrixware's partners and customers as well as to participants in Matrixware or IRF-supported projects. If you have received this documentation in error, please delete the original and any copy or printout thereof.*

---

Points of interest regarding how the data is assembled content-wise.

### **Each publication is extracted as a separate xml instance.**

As legal and technical representations of inventions, patent publications can be grouped in "families" of varying definitions. A domestic family may include a published application, a granted patent, amended versions of either application or patent, or both. It might be expanded to include continuations or divisions of the original application. A "simple" family includes all publications sharing identical priorities (the Espacenet model). An "Inpadoc" Family unites recursively all publications sharing a priority with any other family member.

The Alexandria extract retains each publication (the published application, the search report, the granted patent, the amended patent, etc.) as a separate xml record. Applications using the extract are able to create groupings of patents into families using the simple family ID that is provided for each publication or by manipulating the normalized priority data to build a defined patent family.

### **Multiple formats are provided for applicant and inventor names.**

When available, three sources for applicant and inventor names are provided. Two of these sources come from the Docdb file, and they are labeled as "epo" and "intermediate". "Epo" format contains the "standardized" form of applicant name, while the "intermediate" form follows basic rules for common abbreviations. In both cases, names are capitalized.

The third source for names is "original," and this source is available for those authorities for which we receive data directly from the publishing source. Those authorities include EP, US, WO, JP, DE. The "original" format is the name 'as filed,' and it often includes address information.

### **Data is retained as received; data corrections are logged**

details needed

### **Text fields have language tags when possible**

Titles, abstracts, description sections and claim sections include language labels to enable language-specific queries. Language is inferred when not delivered in original data, and when it cannot be inferred, the language tag is set to a default "xx".

### **Classifications are regularly updated**

IPC Classification (IPCR) is updated by regular weekly updates to the Docdb file, as well as by occasional bulk MCD updates provided by the EPO.

US Class updates are issued 6 times per year by the USPTO, and are applied as received.

### **Legal Status and other Post-Issuance**

details needed

## **Citations (patent and non-patent) are associated with specific publication stages depending on publishing authority**

Excerpt from DocDB Exchange Format, Annex XI

"Country Coverage of the REFI File

Citations from all searches made at the EPO (for BE, CH, EP, FR, GB, NL, TR) and citations for all WO publications are being supported by the master database as well as data from a number of other countries.

*Note that for certain reasons some data may be present only from a given date onwards or data may be incomplete. If you have a very important document it may be worthwhile to check whether references are available, otherwise look them up in esp@cenet by viewing the "original document".*

### **I--Citations in EP and WO records**

From the Docdb Exchange Format Spec:

"Note that for EP and WO, all citations are linked to the first publication of the document. E.g., a search report cited in the EP A3 publication will be linked to the corresponding EP A2."

#### **EuroPCTs:**

1. For EuroPCTs, all citations from the international search report are presented in the first publication of the WO record (and are not repeated in the EP a-level record).

examples:

EP0900000A1 (no pat cites)/WO199835547A1 (3 pat cites)

EP1063901A2 (no pat cites)/WO1999048384A2 (7 pat cites)

EP0426852A1 (3 pat cites)/A4 (no cites)/B1 (no cites)/B2 (no cites)/WO1991015099 (2 cites)

2. If a supplementary search is done by the EPO, citations from that report are presented in the first EP A record

example:

EP0699028A1 (2 pat cites from EPO search rpt.);WO1994026102 (3 pat cites from int'l search rpt.)

?5 additional pat cites are present for these 2 publications: these references were converted from NPL cites to PAJ and WPI into pat cites as a value-add.

3. Citations during examination are presented in the EP A document

example:

EP1200000A1(1 pat cite)/WO2001003597A1(6 pat cites)

#### **European national filing:**

1. Citations from search and opposition are all reported only in the first publication

example:

ep1120451 (A1, B1, and B2)?14 patent citations are present in the A1 record (6 tagged as SEA and 8 tagged as OPP)

The corresponding Register data for patent citations:

Documents cited:

Search

X EP0966955

X WO9900470

XD US4429097

X EP0595590

A US5648326

AD US5916967

Opposition

EP0216416

WO0021493

US5889092

EP1013264

US4842757

EP1090631

WO9800494

US5057241

**II--Citations in US records**